

Topic Analysis through Streamgraph via Shiny Application: a Social Collaborative Approach

Olga Scrivner
CNS, Indiana University
obscrivn@indiana.edu

Vinita Chakilam
Indiana University
vinichak@uemail.iu.edu

Nilima Sahoo
Indiana University
nsahoo@iu.edu

Stephan De Spiegeleire
HCSS
stephandespiegeleire@hcss.nl

Abstract

With the increasing complexity and volume of data, the transformation from streaming information into actionable knowledge becomes more and more challenging and requires a synthesis of computational and substantive approaches. In this view, the collaboration between developers and substantive experts is essential for obtaining meaningful and strategic insights. Despite the large number of various platforms and software to develop a customized tool, the main challenge is developing social organizational forms for communication. In this paper, we explore a new method of organization workflow, namely a social collaboration via the rizzoma platform. In particular, we introduce our on-going project for developing a research-driven visualization portal that is responsive to the need of specific research in strategic studies.

1. Introduction

The information age has transformed data into knowledge that is essential for the well-being of the society and its military and economic strength [1]. In policy making strategic and economic planning, more emphasis has now been placed on data analytics and text mining applications. As Aggarwal and Zhai point out, text mining goes “beyond information access to further help users analyze and digest information and facilitate decision making” [2]. Furthermore, insights from data and strategic predictions become vital and hold “unimaginable potential social consequences of development” for science and technology [3].

On the other hand, the variety and complexity of data makes “the transformation from data to actionable knowledge” more and more challenging [4]. It is often unfeasible to customize the existing tools for specific research needs. Similarly, research scientists are often not equipped with the necessary programming skills or the time to develop their own research-driven tools. Finally, there still exists “a wide chasm between

the general computing community and the scientific computing community” [5]. Given the exploratory nature of research, the development of research software poses challenges to non-scientific developers, as the up-front software specifications are hard to construct [5]. In addition, researchers mainly serve as end-users – a situation that may lead to a developer-user paradox in which “the observing behavior is pointless unless information is presented to the user in a way that meets his or her needs” [6].

In recent years, there has been increasing number of efforts to establish collaboration between software developers and scientists [7, 8]. For example, SBGrid,¹ a consortium of scientific software developers for biology research, functions as an active intermediary between developers and biologists to foster new computational resources. Furthermore, in the field of visualization software, there has been growing interest for interactive social collaboration. In this view, social interaction becomes a part of visualization design [9, 10]. This collaboration is mainly constructed to improve analytical interpretation and perception [11, 10], as the process of mapping from information to visual insights [12] has been hindered by the complexities of software design and non-interactive collaboration approach. As it has been acknowledged, the design of collaborative visualization remains *a grand challenge* for visualization research [13]. With the recent development in interactive web applications (e.g. Shiny) such collaboration becomes feasible not only on a visualization design level but also at the level of functionalities and knowledge mining methods.

In this paper, we introduce our on-going project on a social-scientific collaboration by means of a social platform *Rizzoma* and Shiny, an interactive web framework. In particular, we demonstrate our organizational workflow to build a research-driven tool that is responsive to the need of specific research efforts for strategic studies.² As a part of our collaborative

¹www.sbgrid.org

²<http://hcss.nl/>

design, we develop a user-friendly application for text mining and visual analysis to provide insights on the current social phenomenon of populism. In contrast to the traditional iterative approach that mainly aims to improve functionalities and usability [14], our method mutually enriches the developers and researchers (a.k.a. end-users), allowing for new discoveries and technical creativity.

This paper is organized as follows. In section 2 we introduce the concept of a social collaboration. In particular, we will make use of a philosophical notion *rhizomatic model*. We will also describe the use of streamgraph and its application to topic analysis, a result of our collaborative design. In section 3 we will provide our current workflow that has been developed to improve collaboration and development at all stages of project. Section 4 will present our current development. Finally, we will outline our future direction in section 5.

2. Design

In this section, we will introduce the notion of a social collaboration and describe its potential by illustrating our collaborative design for visualization functionalities.

2.1. Social Collaboration

The growth and complexity of data pose increasing challenges for the research community, as *making sense* becomes more and more challenging [15]. The need for interdisciplinary collaboration has long been recognized [16]. Such collaborations become vital “whenever researchers wish to take their research programs in new directions” [17]. Based on the integration extent, research collaboration could vary from a very low integration, *Investigator-Initiated Research*, to a highly integrated research team, as illustrated in Figure 1 [18].

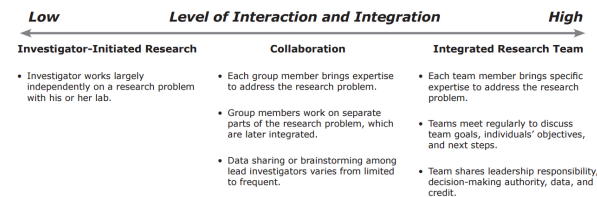


Figure 1. NIH schema: research collaboration

In the field of knowledge visualization, collaboration has also been recognized as an essential component to “create, integrate and apply knowledge” [19]. Synchronous and asynchronous team collaborations have been applied to improve interpretation: for

example, by means of shared displays and remote communication [11, 10]. In the area of commercial software, developers communicate with users mainly to enhance design or functionality [20]. Users, however, are often disengaged from the software development process.

In recent years, there has been growing number of cloud-based applications, ranging from sharing documents on Google drive or Overleaf to real-time code collaboration with Google Hangout. These platforms enable synchronous and remote communication, facilitating team collaboration at all levels of collaboration workflow. In particular, these tools are well designed for a linear workflow (e.g., report, meeting, paper). However, these platforms are not successful when the collaboration requires non-linear communication: for example, for creating new concepts and brainstorming ideas. A cognitive concept mapping [21] is an illustration of this non-linear representation, in which various entries of information are mapped via cross-linking. The disadvantage of cognitive concept mapping lies in its scope, namely mapping new ideas around a given concept. In order to facilitate the workflow for designing a collaborative research-driven tool, multiple concept entries must be also mapped to research domain and various stages of workflow. In what follows, we will describe a philosophical model, known as the *rhizomatic model*, that allows for such collaborative mapping.

The notion of the rhizomatic model has been first introduced in the philosophical essay by Deleuze (1987) as a new social model [22]. In this view, the *rhizomatic* research is viewed as a map of ideas, following principles of multiplicity and heterogeneity. Until recently, the integration of these principles into a research collaboration was not feasible. *Rizzoma* is a recently developed social collaborative platform allowing for such mapping.³ *Rizzoma* is built as a knowledge-management and discussion platform allowing for real-time team communication, planning, brain-storming, and multimedia sharing. This platform allows for continuous project development and team communication as well as concept creation. Figure 2 illustrates a sample of our collaborative project structure, where ideas are organized around concepts and tasks, while preserving a fluidity of conversation among team collaborators, similar to social platforms.

³<https://rizzoma.com>

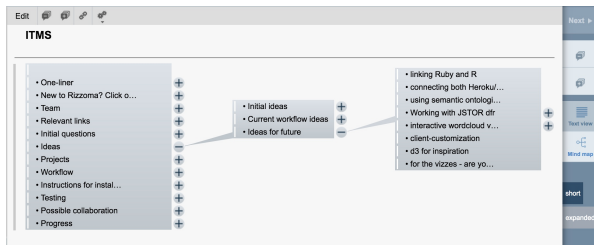


Figure 2. A collaborative platform Rizzoma

2.2. Visualization Design

With large volumes of data, it “becomes more difficult to find and discover what we are looking for” [23]. Visual analytics is a key to unveiling and transforming hidden information into actionable knowledge by facilitating analytical reasoning [13]. In the past, the use of visual tools in a research collaboration had several limitations. First, many applications required programming skills and some were limited in their scope (e.g., visualization type and document format). Second, traditionally researchers remained as end-users often disengaged from the application development. As a result, the choices of functionalities were not always a good fit for their research. In this paper, we demonstrate the advantages of developing applications in direct collaboration between developers and researchers.

Our previous text mining application ITMS (explained in more detail in section 4) was developed in a traditional way with scientists being end-users. The tool offers a variety of useful analyses and visualization types, such as topic modeling, clustering, and word frequency. For example, Figure 3 exhibits a table representing a topic analysis and Figure 4 displays a frequency analysis. Despite the variety of functions and preprocessing steps, the visual analysis of our tool still remains very simple, static, and limited to the overall word distribution.

V1	V2	V3
judicial	courts	law
administrative	case	common
review	eu	act

Figure 3. Topic representation as a table



Figure 4. Cloud representation of word frequencies

To improve its functionality and usability, we have conducted several workshops in the past and received evaluations from the end-users. Despite a number of users’ suggestions, their implementation was seen as unfeasible without a more direct collaboration.

In what follows we will describe the advantages of direct social-scientific collaboration by illustrating how a new interactive visualization tool has been created and built as a result of our collaborative design. The combination of research needs to segment documents by keywords, the interactivity of Shiny applications and social collaboration has led to a new functionality, namely topic analysis from specific segmented documents via a dynamic streamgraph displaying the change of topics (a.k.a main ideas) over time.

2.3. Streamgraph visualization

Streamgraph, which is a variation of stacked area graph, was developed in 2008 to visualize trends in personal music listening [24]. It was then noticed that the streamgraph was “capable of conveying a large amount of data in a manner that engages mass audiences” [24]. Visual interpretation of data was facilitated by the use of size and colors for individual stream (a.k.a category) [25]. As the change in data over time was shown in a stream-like shape, this graph became known as a streamgraph.

Several applications have been developed to represent data as a streamgraph. The authors Byron and Wattenburg open-sourced their implementation of streamgraph in Java and *Processing*.⁴ Furthermore, streamgraph is enabled in *Paper machines*,⁵ a plugin to the bibliographic management software Zotero. This web-based plugin is designed to support researchers

⁴<https://github.com/leebyron/streamgraph>

⁵<http://papermachines.org/>

by providing user-friendly access to visualization and analysis that otherwise would require programming skills. To illustrate Paper machine visualization, we have taken the following examples from a Zotero corpus. Figure 5 shows the distribution of topics over time. This representation is a traditional streamgraph function centered in the y-axis and with smoothed *streams*.

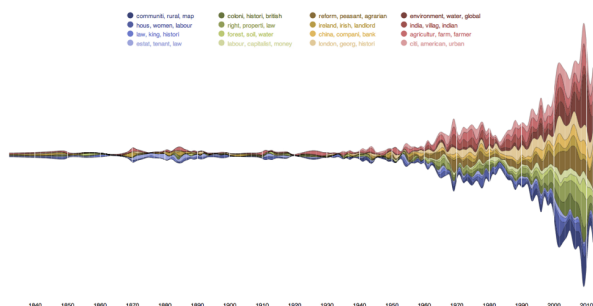


Figure 5. Streamgraph and topic analysis: Paper Machine - Example created by Chris Johnson-Roberson (Creative Commons Attribute 3.0 Unported)

Figure 6 illustrates a variation of streamgraph, where the y-axis is extended. In addition, the interactive display enables the access to individual word frequencies.

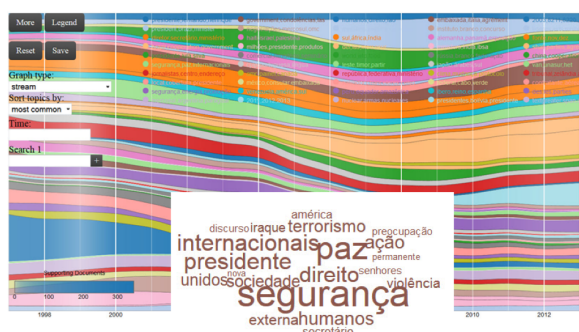


Figure 6. Streamgraph and word cloud: Paper Machine [26]

Streamgraph also became available in the R programming language and environment. The streamgraph R package⁶ is an htmlwidget that is based on the D3.js JavaScript library. With this library, streamgraph allows for interactive components, such as following *flow* or filtering the view. In contrast to Java implementation, R is built for data manipulation

⁶<https://github.com/hrbrmstr/streamgraph>

and visualization, which makes it a good choice for streamgraph visualization. In addition, R has a Shiny library that facilitates the creation of interactive web interfaces. Thus, R seems to be a good choice for our collaborative project, providing interactivity, flexibility for customization, and a large library of available functions.

3. Project workflow

In our collaborative design we follow the subsequent notions developed in our recent work [27]:

1. *Learning process* [28] - this process includes conversations between scientists and developers about research questions and available computational solutions (e.g., text mining methods and visual analysis). This stage is designed to provide mutual learning experiences: scientists learn about software development and developers acquire knowledge about research.
2. *Collaborative process* - this process includes the exchange of ideas, tasks creation, and feedback. Scientists assess the quality of suggested methods with real data, whereas developers assess the feasibility of research questions and available computational methods.

It should be noted that these processes are iterative and non-hierarchical, as the learning process also occurs during the collaborative process. In this view, the project schema represents a constant non-linear flow from ideas to design and development via *Rizzoma*, illustrated in Figure 7.

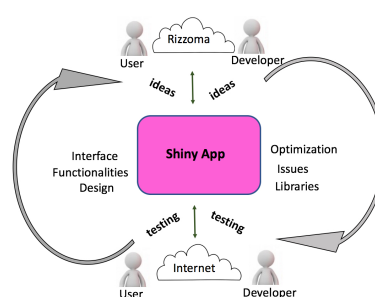


Figure 7. Workflow: iterative design, integration, and collaboration

The research development consists of three main stages:

1. *Ingestion*: Selection, organization, and storage of data

2. *Digestion*: The processing and analysis of information

3. *Egestion*: Production of insightful output

In all three stages, collaboration is enabled via *Rizzoma*⁷ and an interactive Shiny application.⁸ Shiny is a web application allowing for interactive data manipulation, analysis and visualization. Built with R as a back-end, Shiny web framework also provides access to advanced text mining and quantitative algorithms. There are several advantages of using Shiny application for data science and data visualization. First, the Shiny application offers a user-friendly graphical interface. Second, there is no need for local installation, as the deployed application is web-based and accessible from any browser, including mobile devices. Third, the development of the graphical user interface only requires knowledge of the R programming language. It is flexible and efficient, thus allowing for quick modification and adjustment during the development stage. Finally, R is built for data analytics and visualization, thus providing access to a large number of available libraries and state-of-the-art statistical methods.

These two tools enable an effective interactive workflow: First, the researcher and the developers create nodes for *team*, *initial questions*, *ideas*, *relevant links*, and *one-liner*, a one-sentence summary of the project, as illustrated in Figure 2 (section 2.1). Various types of brainstorming and communication are then initiated through mentioning (similar to google doc), task assignment, and interlinking with other nodes. Second, the demos are embedded as Shiny web frames in *Rizzoma*, so that the researcher can evaluate and provide feedback without any need for downloading demos. In addition, data and code are also shared via interlinks between the researcher and developers.

4. Interactive text Mining Suite

In our previous work [29, 30], we laid a foundation for a developing interactive analytical applications. Using Shiny web framework, we have developed a prototype, *Interactive Text Mining Suite* (ITMS),⁹ designed to assist researchers to explore and visualize heterogeneous data without any need for programming or software downloading. This prototype is built using R as a back-end and Shiny app as a front-end, which allows for the creation of an interactive and user-friendly interface. Our choice of R and Shiny

is directly motivated by the analytical strength of R, which is often referred to as “the lingua franca of data science”.¹⁰ R is not just a programming language; it is rather a software for data analysis and visualization. Furthermore, the recently developed Shiny framework makes it possible to merge the analytical power of R with web interactivity, thus helping to increase the literacy in data analytics among non-programmers and provide tools for data-driven research in social and political science. The ITMS is currently hosted on the *shinyapps.io* platform, a cloud service operated by RStudio.

Based on *Rizzoma* collaboration, we have designed ITMS functionalities to fit the need of the research on strategic insights. The data are based on a Zotero library collection (Figure 8).


- [EBSCO](#) (193)
- [IngentaConnect](#) (200)
- [Science Direct](#) (33)
- [Scopus](#)(41)
- [Springer](#)(28)
- [Springer](#)(66)
- [Web of Science](#)(10)
- [Wiley](#)(62) 

Figure 8. Zotero corpus description in *Rizzoma*

Despite a number of interactive visualizations provided by *Paper Machine* (see section 2.3), the current research by HCSS scientists requires a more customized text preprocessing, namely a text segmentation based on the search for certain keywords. In addition, the window context, namely the number of words around the search terms, has to be customizable. The following examples demonstrate various types of queries that have been developed during the exploratory research stage. Sometimes the left and right context specifications are required, which is represented by an integer next to the search. This task is a good illustration of the advantages of our social-scientific collaboration. These queries were not specified initially, they came across as the researcher began exploratory analysis, which would pose challenges for a traditional software development.

- populism AND (defence OR defense OR military)
- populist AND (defence OR defense OR military)
- populist influence on (defence OR defense OR military) 100

⁷<http://rizzoma.com/>

⁸<https://www.shinyapps.io/>

⁹<http://www.interactivetextminingsuite.com/>

¹⁰<http://blog.revolutionanalytics.com/2013/11/the-rise-of-r-as-the-language-of-analytics.html>

- “role of populism” AND (defence OR defense OR military) 10
- populism AND influence AND (defence OR defense OR military) 10
- “populism defense” 100

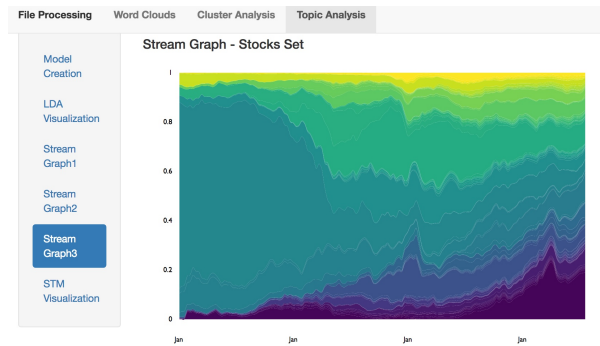


Figure 10. Streamgraph prototype II: Shiny

The Shiny web application allows for an instant evaluation of text segmentation by a scientist and a developer at the same time. In addition, Shiny Reactive Log Visualization function provides an interactive summary of execution. We plan on using this function to optimize our pre-processing and text-mining sequences (e.g., order of steps and most actively used functions or visualizations).

Currently, we are also testing the visual interpretation of streamgraph for topic analysis.¹¹ The implementation of R package *stm* provides a proportion of topics over time instead of count numbers used in the original streamgraph design.¹² Figure 9 and 10 display various types and options for display, which are discussed in *Rizzoma*. The design and discussions about visualization types are another mutual enrichment during this social-scientific collaboration, as we develop and test research-driven visualization.

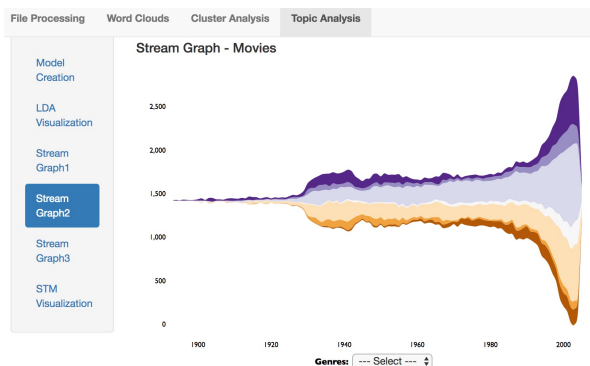


Figure 9. Streamgraph prototype I: Shiny

5. Conclusion and Future Directions

The need for collaborative research has long been acknowledged. Until recently, however, synchronous collaboration was possible only using a linear communication of ideas. As the knowledge is typically built around concepts, concept mapping seems to fit and motivate collaborative learning and development. In our project we have applied such mapping using a web-based platform *Rizzoma*. The *Rizzoma* model provides us with a space for creativity while maintaining a workflow. Shiny web framework is used to interactively evaluate and test the implementation of the ideas and tasks from *Rizzoma* workflow. By embedding the Shiny application demos within the *Rizzoma* platform, we are also able to maintain our collaboration in the same social environment. As a result, these tools made it possible to design and create a new functionality, namely topic modeling via streamgraph from segmented contexts.

Furthermore, this research also contributes to the field of research-driven collaborative design and visualization. While most visualization research has explored the cognitive and perceptual aspects of design, social interaction has only recently been recognized as a part of visualization system design. In this view, the choice of functionalities is contingent upon direct engagement and collaboration between tool-developers and researchers.

In the future, we plan on evaluating the social-scientific collaboration among graduate students as developers. Typical student collaborations involve a linear communication via Google doc and various social media channels. We would like to investigate whether the introduction of non-linear workflow will encourage students' creativity and collaboration.

Finally, we would like to optimize our functionalities and preprocessing steps, based on the Shiny logs and further evaluation.

¹¹<https://languagevariationsuite.shinyapps.io/TextMiningZotero/>

¹²<https://github.com/obsrvin/HCSS-Rizzoma>

Acknowledgment

We would like to thank Todd Theriault for editing an earlier draft of this paper. This work was partially funded by the National Science Foundation under grant 1713567. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- [1] P. Drucker, "The age of social transformation," *Atlantic Monthly*, vol. 274, pp. 53–80, 1994.
- [2] C. C. Aggarwal and C. Zhai, "An Introduction to Text Mining," in *Mining Text Data*, pp. 1–10, Boston, MA: Springer US, 2012.
- [3] K. Börner, *Atlas of Knowledge: Anyone Can Map*. 2015.
- [4] L. Cao, *Metasynthetic Computing and Engineering of Complex Systems*. London: Springer-Verlag, 2015.
- [5] J. E. Hannay, C. MacLeod, J. Singer, H. P. Langtangen, D. Pfahl, and G. Wilson, "How do scientists develop and use scientific software?," in *2009 ICSE Workshop on Software Engineering for Computational Science and Engineering*, pp. 1–8, IEEE, may 2009.
- [6] C. Jeffery and J. Al-Gharaibeh, *Writing virtual environments for software visualization*. 2015.
- [7] A. Morin, B. Eisenbraun, J. Key, P. C. Sanschagrin, M. A. Timony, M. Ottaviano, and P. Sliz, "Cutting Edge: Collaboration gets the most out of software," *eLife*, vol. 2, p. e01456, 2013.
- [8] T. A. Finholt, "Collaboratories as a new form of scientific organization," *Economics of Innovation and New Technology*, vol. 12, pp. 5–25, jan 2003.
- [9] F. Viégas and M. Wattenberg, "Communication-minded visualization: A call to action," *IBM Systems Journal*, vol. 45, 2006.
- [10] J. Heer and M. Agrawala, "Design Considerations for Collaborative Visual Analytics," *Information Visualization*, vol. 7, pp. 49–62, 2008.
- [11] K. Brodlie, D. Duce, J. Gallop, J. Walton, and J. Wood, "Distributed and collaborative visualization," *Computer Graphics Forum*, vol. 23, pp. 223–251, 2004.
- [12] M. O. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization : Foundations, Techniques, and Applications*. CRC Press, 2010.
- [13] J. Thomas and K. Cook, eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Press, 2005.
- [14] J. Nielsen, "Iterative User-Interface Design," *Computer*, vol. 26, no. 11, pp. 32–41, 1993.
- [15] S. Negru and S. C. Buraga, "An educational tool for an interactive faceted exploration of dbpedia life sciences data," in *Web Information Systems Engineering - WISE 2013 - 14th International Conference, Nanjing, China, October 13-15, 2013, Proceedings, Part II*, pp. 503–506, 2013.
- [16] X. Sigma, "Removing the boundaries: Perspectives on cross-disciplinary research: Final report on an inquiry into cross-disciplinary science," 1998.
- [17] F. L. Macrina, "Dynamic Issues in Scientific Integrity: Collaborative Research," tech. rep., Washington, DC, 1995.
- [18] M. L. Bennett, H. Gadlin, and S. Levine-Finley, "Collaboration and Team Science: A Field Guide - Team Science Toolkit," tech. rep., 2010.
- [19] M. J. Eppler, "What makes an effective knowledge visualization? a review of seminal concepts," in *Proceedings of the 15th International Conference on Information Visualization. IEEE*, 2011.
- [20] J. Whitehead, "Collaboration in Software Engineering: A Roadmap," in *2007 Future of Software Engineering*, (Washington, DC, USA), pp. 214–225, IEEE Computer Society, 2007.
- [21] J. D. Novak, "Concept maps and Vee diagrams: two metacognitive tools to facilitate meaningful learning," *Instructional Science*, vol. 19, no. 1, pp. 29–52, 1990.
- [22] G. Deleuze and F. Guattari, *A thousand plateaus : capitalism and schizophrenia*. University of Minnesota Press, 1987.
- [23] D. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.
- [24] L. Byron and M. Wattenberg, "Stacked Graphs - Geometry and Aesthetics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1245–1252, 2008.
- [25] J. Mackinlay, "Automating the design of graphical presentations of relational information," *ACM Transactions on Graphics*, vol. 5, no. 2, pp. 110–141, 1986.
- [26] S. D. Spiegeleire, E. Chivot, O. Andriana, M. Demirel, V. Miladinova, M. Mukherjee, J. Silveira, M. Y. Yang, and O. Zelinska, *What the Official Websites Say: Developing and Testing a New Systematic Information Collection Method*. The Hague Centre for Strategic Studies, July 2014. 00003.
- [27] O. Scrivner, V. Chaklam, J. Poojary, N. Sahoo, C. Uppuluri, and S. De Spiegeleire, "Building Customized Text Mining Tools via Shiny Framework: The Future of Data Visualization," in *The 28th Modern Artificial Intelligence and Cognitive Science Conference, At Fort Wayne, Indiana*, 2017.
- [28] P. Dillenbourg, *Collaborative learning : cognitive and computational approaches*. Pergamon, 1999.
- [29] O. Scrivner and J. Davis, "Topic Modeling of Scholarly Articles: Interactive Text Mining Suite," in *Computational Linguistics and Intellectual Technologies: International Conference "Dialogue 2016"*, 2016.
- [30] O. Scrivner and J. Davis, "Interactive Text Mining Suite: Data Visualization for Literary Studies," in *the Workshop on Corpora in the Digital Humanities*, pp. 29–38, 2017.